Revista Digital Universitaria
10 de julio 2008 • Volumen 9 Número 7 • ISSN: 1067-6079

ARTICULO

LA ERA DE LAS MÁQUINAS LECTORAS

José Antonio Millán (http://jamillan.com) es lingüista especializado en el campo de la edición (con especial atención a la digital) y la cultura escrita. Mantiene un blog sobre estos temas: http://elfuturodellibro.com.

10 de julio 2008 • Volumen 9 Número 7 • ISSN: 1067-6079

José Antonio Millán (http://jamillan.com) es lingüista especializado en el campo de la edición (con especial atención a la digital) y la cultura escrita. Mantiene un blog sobre estos temas: http://elfuturodellibro.com.

Este texto comenzó como una conferencia en el Seminario Litterae (Madrid) de septiembre del 2007. Gracias a Antonio Castillo, Vanessa de Cruz y Emilio Torné por su invitación a participar. Gracias a Javier Candeira por su ayuda. Una segunda presentación oral tuvo lugar en junio del 2008 en el 4° Foro de Edición Digital (México D.F.); gracias a Ernesto Priani por la invitación. Una versión ampliada de este texto aparecerá en la revista Arbor http://arbor.revistas.csic.es/.

"http://www.revista.unam.mx/vol.9/num7/art52/int52.htm

INICIO

Un fantasma recorre el universo de los textos. Un ejército de máquinas, a las que aludimos con metáforas zoológicas (ar*añas*) o mecánicas (*cosechadoras*), merodean por la Red, leen nuestros textos, e incluso atisban por encima del hombro mientras escribimos.

¿Paraquélohacen?¿Paraespiarnos?Aveces...¿Paracomprendernosmejor?Ciertamente.¿Paraayudarnos?Eso dicen...

En el universo de la World Wide Web las máquinas (los ordenadores, o mejor dicho, sus programas) saltan constantementedepáginaenpáginaatravésdelosenlaces,escudriñansucontenidoyalmacenancadapalabra ycadacombinación.Deesaforma,cuandolespreguntamos:¿dóndesehabladeHércules?,puedencontestarnos: *aquíyallá*...Perotambiénleenlosenlaces,yasíseenterandequécreenlosautores(depáginasweb,decualquier documento accesible en la Red) que tratan las páginas a donde remiten....

Precisamente esa lectura de enlaces es la responsable de algunos de los hallazgos más asombros de los buscadores: encontrar lo que no está... Por ejemplo, la búsqueda de *gentuza* en Google lleva a esta noticia (que no contiene la palabra en cuestión):

A esta mujer, víctima del Katrina, la han dejado sin un duro de la ayuda que recibió como afectada por el huracán. Todo fue porque le hicieron una foto en primer plano con su tarjeta en la mano. Al poco tiempo de publicarse en diversos medios digitales la instantánea de la Agencia France Press, realizaron una serie de compras en Internet con su número de Mastercard.

II

Ocasionalmente,lasmáquinastambiénescriben(o,paranoexagerar:editan,ponenencontactotextosdiversos). Ocurre,porejemplo,cuandocolocandentrodelaspáginaswebanunciosrelacionadosconsutema(queeslo que hace Google Adwords[1]).

Paraellotienenquehaberleídosucontenido.Porejemplo,enunapáginaqueanalizaunoscartelesamenazadores[2] aparecen estos anuncios:

> El centro del accidentado. Ayuda jurídica para víctimas de accidentes.
> Chistes de abogados
> Problemas con alquileres

¿Por qué? El texto contenía términos como am*enaza, insulto, violencia, transgresor o merodeador*, junto a expresiones como "me cago en sus muertos". Los insondables algoritmos de Google Adwords han determinado que (entre los temas de publicidad que administran) los relacionados con *accidentes, abogados y problemas* eran los más pertinentes...

1 http://adwords.google.com/
2 http://jamillan.com/florarma.htm

10 de julio 2008 • Volumen 9 Número 7 • ISSN: 1067-6079

Máquinas que entienden

Este tipo de comportamientos nos podría llevar a la siguiente cuestión. Sí: las máquinas leen nuestras páginas web, pero, ¿las entienden? En realidad, esto es una variante del Test de Turing. Como se recordará, en dicha prueba un humano conectado a un terminal exclusivamente textual (tipo chat) debe determinar, sólo a través del diálogo, si al otro lado hay una máquina o un ser humano.

Uno dice "¡Gentuza!", y el buscador contesta: "Sí, como esos que estafaron a una víctima del Katrina...". Uno escribe "amenaza, violencia, transgresor", y los anuncios corean: "abogados, accidentes, problemas". ¿No se están entendiendo las máquinas? Bueno: lo suficiente como para echarnos una mano. Y el éxito de los buscadores y de los programas de anuncios contextuales parece indicar que lo logran...

Hay en marcha sistemas todavía más sofisticados. Por ejemplo: un programa que analiza, en un foro sobre valores bursátiles, cuál es la opinión generalizada sobre cuáles van a subir y cuáles a bajar. Es el Community Sentiment de Yahoo[3]. Un análisis de este estilo exige manejar un número considerable de variables semánticas y pragmáticas.

III

Y en este momento nos surge un tema de especial interés. Si las máquinas nos leen, ¿no habrá que tenerlas en cuenta cuando escribimos? La respuesta es claramente que sí: el autor o editor de cualquier material en la Web tiene que favorecer que le lean las máquinas, so pena de comprometer su propia difusión.

Un ejemplo particularmente ilustrativo es el de las licencias Creative Commons[4]. Cada una de ellas tiene tres versiones:

el resumen, legible por humanos. Dice cosas como:

> Usted es libre de: copiar, distribuir y comunicar públicamente la obra.

el código legal, legible por abogados; éste es su comienzo:

> LA OBRA (SEGÚN SE DEFINE MÁS ADELANTE) SE PROPORCIONA BAJO LOS TÉRMINOS DE ESTA LICENCIA PÚBLICA DE CREATIVE COMMONS ("CCPL" O "LICENCIA"). LA OBRA SE ENCUENTRA PROTEGIDA POR LA LEY ESPAÑOLA DE PROPIEDAD INTELECTUAL Y/O CUALESQUIERA OTRAS NORMAS RESULTEN DE APLICACIÓN. QUEDA PROHIBIDO CUALQUIER USO DE LA OBRA DIFERENTE A LO AUTORIZADO BAJO ESTA LICENCIA O LO DISPUESTO EN LAS LEYES DE PROPIEDAD INTELECTUAL.

el código digital, legible por máquinas

> <rdf:RDF xmlns="http://creativecommons.org/ns#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"><License rdf:about="http://creativecommons.org/licenses/by/2.5/es/"><permits rdf:resource="http://creativecommons.org/ns#Reproduction"/><permits rdf:resource="http://creativecommons.org/ns#Distribution"/><requires rdf:resource="http://creativecommons.org/ns#Notice"/><requires rdf:resource="http://creativecommons.org/ns#Attribution"/><permits rdf:resource="http://creativecommons.org/ns#DerivativeWorks"/></License></rdf:RDF>

3 http://finance.yahoo.com/newfp.
4 http://creativecommons.org

"http://www.revista.unam.mx/vol.9/num7/art52/int52.htm

Este caso reflejabien las complejidades de la autoría/edición en la Web. Una norma general de redacción es tener en cuenta al destinatario de un texto para ajustar su nivel, y eso justifica la diferencia entre el "código legal" y el "resumen": sus receptores son, respectivamente, el abogado y el lego en la materia.

El código digital está destinado a ser leído por sistemas automáticos. En el caso de Creative Commons, se ha incluido para informar a los buscadores que quieran localizar contenidos con determinados tipos de licencia. Las personas no tienen por qué entenderlo, y ni siquiera leerlo: el texto no está visible en la página.

Y una última, pero importante consecuencia, para aquellos que escriben o editan en la Web: cada enlace es un voto a una página. Y mediante el texto específico que enlazamos estamos diciendo algo sobre la página de destino no sólo a nuestros lectores humanos, sino, sobre todo, a las máquinas.

Máquinas que ayudan

IV

Además de las búsquedas, que antes veíamos, las máquinas también están leyéndonos para ayudarnos con distintas tareas...

Los servicios de alertas, como Yahoo Alerts[5], rastrean la prensa y otras páginas web para avisarnos de cuándo aparece alguna de las palabras clave que les hemos indicado. Resulta muy útil para tener controlada a una empresa rival, conocer los movimientos de una determinada persona, o sencillamente, ver qué dicen de nosotros (ego surfing).

Los detectores de plagios, como Damocles[6], comparan el texto que les sometamos con muchos otros dispersos por la Web, con el objeto de determinar si se han utilizado (sin citar) partes de otras obras.

Los sintetizadores de voz (como SodelsCot[7]) leen los textos que les proponemos.

Sin olvidar a los programas traductores (como SoftCatalà[8], del catalán al castellano y viceversa), que leen nuestros textos para traducirlos.

Y por último, el sistema de espionaje anglosajón ECHELON (gobernado por Estados Unidos, Canadá, Gran Bretaña, Australia, y Nueva Zelanda) o el sistema Carnivore del gobierno de los Estados Unidos (FBI) escrutan las comunicaciones (correos electrónicos, por ejemplo) a la búsqueda de términos o nombres. Lo bajo de sus fines no debe hacernos olvidar la magnitud de la tarea que afrontan.

V

Hasta aquí nos hemos movido en un dominio, el digital, que posibilita que las máquinas nos lean directamente. En la página web los humanos vemos formas, descifra mos signos y por último leemos palabras. Las máquinas también las leen, pero no por el dibujo que pintan en la pantalla (el cual puede cambiar según las preferencias de nuestro navegador), sino porque acceden al código que les representa. Por ejemplo: la H tiene el código hexadecimal 48, y el fragmento de código

%48%E9%72%63%75%6C%65%73

5 http://alerts.yahoo.com.
6 http://viper.csse.monash.edu.au/damocles/about/
7 http://www.sodels.com/
8 http://www.softcatala.org/traductor/.

10 de julio 2008 • Volumen 9 Número 7 • ISSN: 1067-6079

se leería *HERCULES*. El sintetizador de voz que lee el documento de procesador de textos y el programa espía que supervisa nuestro correo acceden también al código de las letras.

En caso de contradicción entre el mensaje visual y el código los humanos seguimos, por supuesto, lo que nos dicen nuestros ojos. Por eso en los años 80, para burlar la censura que supervisaba las BBS (tablones de anuncios electrónicos), los usuarios escribían sustituyendo letras por otros signos con los que tenían cierto parecido (pero que no compartían su código). Por ejemplo, para escribir *similar* se usaba la siguiente secuencia de caracteres:

51m1L4R

Lamentablemente, ya hay programas que leen también estas escrituras...

Ancla 4
Máquinas que describen, máquinas que leen

Pero aparte de este acceso directo al código, las máquinas están leyendo cada vez más las publicaciones impresas.

Hay dos formas en que las máquinas pueden tratar nuestros textos impresos. Una es fotografiando sencillamente el texto, es decir, describiendo pixel a pixel la traza de sus letras.

La Fig. 1 muestra las tres letras iniciales de la palabra Hercules en el facsímil JPEG de la primera edición del Quijote en la Biblioteca Virtual Cervantes.

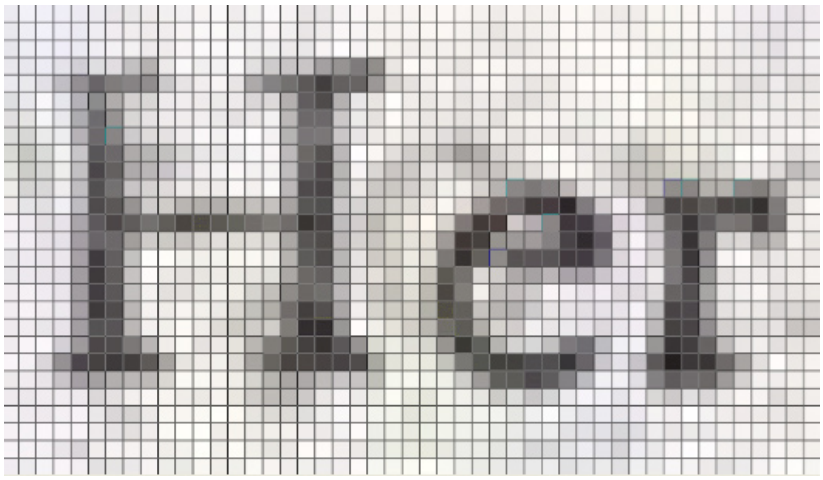

Fig. 1: Descomposición en píxels de un fragmento de imagen con tres letras.

Describir la forma de los signos alfabéticos no es un comportamiento muy sofisticado. Pero el lector interpreta la alineación de píxels (Fig. 2).

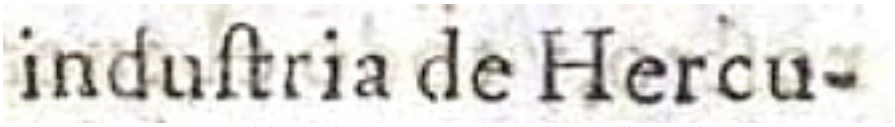

Fig. 2: Palabras tal y como aparecen en la pantalla (ampliación).

Ahora bien, ¿sabe leer el escaneador de páginas ante el que desfilaron las páginas del Quijote? Claramente, no. El portador de formas de letras no lee.

"http://www.revista.unam.mx/vol.9/num7/art52/int52.htm

VII

Para que las máquinas lean de verdad hay que ir un paso más allá: Google Libros[9] (por poner un ejemplo bien conocido) está digitalizando libros de las bibliotecas. Pero además de fotografiar sus páginas les aplica un programa de reconocimiento óptico de caracteres (OCR).

A través de ese procedimiento, la máquina reconocerá la forma que "tiene primero dos líneas, y otra las separa en el centro" como una hache mayúscula (si el texto está en alfabeto latino) o como una eta mayúscula (si está en griego). Y así sucesivamente. Por ejemplo, sometamos el archivo con las palabras del Quijote de la Fig. 2 a un OCR. Nos dará este resultado:

indufhia de Hercu-

Como vemos, puede haber errores. En este caso, la tipografía del XVII tiene ligaduras (como la que une s y t) que el programa no reconoce: en seguida veremos cómo lidiar con ellos. Pero en casos más modernos o claros la máquina puede leer todo el texto satisfactoriamente.

Al final del proceso, el OCR habrá extraído del "cuerpo" del libro (el papel y la tinta) su "alma", el conocimiento de la secuencia de caracteres que lo constituyen: la "acertada disposición del impresor y corrector", en palabras del impresor del XVII Alonso Víctor de Paredes.

El OCR hace que los impresos se fundan en el *continuum* digital del que ya formaban parte las páginas web y otros archivos accesibles por Internet. Y cuando preguntemos: ¿en qué obras se encuentra la palabra "Hércules"?, acudirán a respondernos no sólo las páginas web, sino también las de los libros.

Máquinas que aprenden

VIII

Por último, veamos cómo los humanos estamos, enseñando a las máquinas a perfeccionar su lectura.

Captcha[10] es el sistema mediante el que un sitio web con intervención del público se defiende de los programas que se dedican a introducir spam, proponiendo a los usuarios que tecleen el texto de una secuencia de letras deformada o borrosa que se les ofrece (Fig. 3).

Fig. 3. Captura del captcha de un blog.

9 http://books.google.es/.

10 http://es.wikipedia.org/wiki/Captcha

10 de julio 2008 • Volumen 9 Número 7 • ISSN: 1067-6079

Esta tarea exige (al menos por el momento) un ser humano, y en ese sentido es un test de Turing.

Pues bien: ha nacido reCaptcha[11] (Fig. 4). Su peculiaridad es que el texto que propone para interpretación proviene del escaneado de libros: son palabras que el reconocimiento óptico de caracteres no acierta a interpretar (como industria, que veíamos anteriormente). El programa de OCR detecta una palabra problemática y reCaptcha la ofrece como clave de acceso, emparejada con otra palabra cuya interpretación se conoce (y que sirve de control).

Fig. 4. Recaptcha.

Las palabras dudosas se ofrecen cierto numero de veces, hasta que la lectura se confirma.

ReCaptcha está funcionando por el momento como una ayuda para las digitalizaciones del Open-Access Text Archive[12]. Teniendo en cuenta que cada día se resuelven 60 millones de Captchas, que llevan de media 10 segundos, su suma daría 150.000 horas de trabajo al día, que reCaptcha pondría al servicio de la digitalización de libros.

Comentarios finales

...Y éste es el panorama: ejércitos de autómatas rastreando el ciberespacio y hordas de máquinas leyendo las bibliotecas. Programas que descifran letras y humanos que les ayudan, porque así se ayudan a sí mismos.

Más círculos: humanos que preguntan a la máquinas dónde están las cosas que les interesan, para luego escribir textos que leerán las máquinas para a su vez contarle a otros humanos de qué tratan.

Este espacio simbiótico de personas y máquinas, este *continuum* digital de textos y códigos es el caldo de cultivo de la cultura actual.

11 http://recaptcha.net/learnmore.html
12 http://www.archive.org/details/texts